Harvard University

Harvard University Biostatistics Working Paper Series

Year 2005 *Paper* 22

Feature-Specific Penalized Latent Class Analysis for Genomic Data

E. Andres Houseman* Brent A. Coull†

Rebecca A. Betensky‡

*Harvard School of Public Health, ahousema@hsph.harvard.edu

†Harvard School of Public Health, bcoull@hsph.harvard.edu

‡Harvard School of Public Health, betensky@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper22

Feature-Specific Penalized Latent Class Analysis for Genomic Data (Technical Report)

E. Andrés Houseman
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave., Boston, Massachusetts 02115, U.S.A

Brent A. Coull
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave., Boston, Massachusetts 02115, U.S.A

Rebecca A. Betensky
Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave., Boston, Massachusetts 02115, U.S.A

September 16, 2005

Summary

Genomic data are often characterized by a moderate to large number of categorical variables observed for relatively few subjects. Some of the variables may be missing or noninformative. An example of such data is Loss of Heterozygosity (LOH), a dichotomous variable, observed on a moderate number of genetic markers. We first consider a latent class model where, conditional on unobserved membership in one of k classes, the variables are independent with probabilities determined by a regression model of low dimension q. Using a family of penalties including the ridge and lasso, we extend this model to address higher dimensional problems. Finally, we present an orthogonal map that transforms marker-space to a space of "features" for which the constrained model has better predictive power. We demonstrate these methods on LOH data collected at 19 markers from 93 brain tumor patients. For this data set, the existing unpenalized latent class methodology does not produce estimates. Additionally, we show that posterior classes obtained from this method are associated with survival for these patients.

1 Introduction

A common outcome in genetic sudies is *loss of heterozygosity* (LOH), which refers to the loss of one allele of a chromosomal region in a tumor cell. For one subject at a single genetic marker, LOH, a binary variable, is ascertained by sequencing the marker for both normal and tumor cells obtained from the subject. If the normal cells demonstrate heterozygosity but the tumor cells do not, then loss (LOH = 1) is deemed at the marker location. If both normal and tumor cells demonstrate heterozygosity, then there is no

loss, and LOH = 0. However, if the normal cells demonstrate homozygosity, then LOH is deemed noninformative, or missing from a statistical perspective.

LOH is of interest in cancer studies because it may suggest the loss, through mutation or other cell injury, of a tumor suppression gene. For instance, allelic losses on chromosome 1p have been found frequently in anaplastic oligodendrogliomas, a common variant of brain tumor. Furthermore, the characterization of LOH on chromosomes 1p and 19q is of prognostic interest, since it has been shown to be highly associated with response to chemotherapy and long survival in patients with certain malignant brain tumors (Cairncross *et al.*, 1998; Ino *et al.*, 2001). Previous analyses of LOH in oligodendroglioma used three CA-repeat polymorphism markers to assess LOH of the whole chromosome arm. An entire chromosome arm was assumed to be lost if LOH was observed at all informative markers among those three. Thus, LOH on a given chromosome arm was coded as a binary variable.

Recently, a "medium throughput" quantitative method for assessing LOH at nineteen approximately equally-spaced markers on chromosomes 1p and 19q was applied to a sample of 93 brain tumors. This method was used previously in colorectal tumors (Cawkwell *et al.*, 1993), but formal statistical analysis was not applied. The brain tumors in the current study come from two different sources: one sample consists of subjects from both the London Regional Cancer Centre in London, Ontario and the Massachusetts General Hospital in Boston (termed "MGH"cases) and the second sample consists of subjects from the Henry Ford Hospital in Detroit (termed "HFH" cases). Although nominally of the same diagnosis, the MGH group was evidently more homogeneous in diagnoses and exhibited more LOH than the HFH group. Also, the MGH group was followed about twice as long as the HFH group (median follow-up of 150 months versus 78 months).

One question of interest is whether there is heterogeneity of LOH across markers

covering chromosome 1p, or whether the entire chromosomal arm is typically lost. If there is heterogeneity, it is of interest to identify a few common LOH profiles that define clusters of patients. Ultimately, it is of interest to investigate the association of these profiles with survival. Though not the subject of the present paper, tumors with heterogenous patterns of LOH are potentially informative regarding the location of tumor suppressor genes (Dong *et al.*, 2004).

A typical approach to analyzing genomic data is to conduct a univariate analysis on each genetic marker and adjust inference using techniques that control the familywise error rate or false discovery rate (Westfall and Young, 1993; Benjamini and Hochberg, 1995). However, such approaches are not well-adapted to the setting where variables are highly correlated and each analysis may involve a large fraction of missing variables. Moreover, it is unclear how such methodologies can be adapted to obtain clusters, or profiles, that describe the correlation among variables.

One approach to the ascertainment of LOH-based clusters is latent class analysis. Latent class models are widely recognized as categorical variable analogs to factor analysis (Bartholomew, 1987; Bandeen-Roche *et al.*, 1997). These models postulate the joint distribution of a large number of observed discrete variables as a mixture of only a few distributions, defined by the value of a latent categorical variable. The models cluster the subjects into the unobserved classes according to similarity in pattern of the observed variables, typically under the assumption that the variables are independent given membership in a particular class. One limitation of the latent class approach is the number of conditional probabilities that can be considered without overfitting the data. In particular, if each variable is assigned its own success probability conditional on class membership (an *unrestricted* model), then the number of unknown parameters scales linearly with the number of variables. Certainly, if the number of variables is small

relative to the sample size, then it is feasible to assign each variable its own set of conditional probabilities. However, in high-dimensional settings common to genetic studies, it is often the case that unrestricted latent class models contain too many parameters. In the LOH example, there are 19 LOH variables for each of 37 MGH subjects and 56 HFH subjects. A large percentage of these variables is missing due to homozygosity of normal cells: on average, 5.1 of 19 markers, or 27%. Straightforward application of the methodology proposed by Bandeen-Roche *et al.* (1997), involving at least 59 parameters for a 3-class model, led to a divergent algorithms or Hessians that were not positive definite for every initial value we used.

One could potentially overcome the unwieldy dimension of an unrestricted latent class model by placing constraints on a subset of parameters in the model. However, existing constrained latent class models that have been considered in the literature are too restrictive. Several authors have considered parsimonious extensions of the basic latent class model, obtained by constraining model parameters. Early approaches fixed some of the conditional probabilities of a success, given class membership, to given values or constrained them to be equal (Lazarsfeld and Henry, 1968). Along the same lines, Agresti and Lang (1993) and Lindsay *et al.* (1991) considered models in which the associations between the latent class and the observed variables are the same for all variables. Meulders *et al.* (2002) considered constrained latent class models in which the conditional probabilities are a nonlinear function of a smaller set of basic parameters. Other recent work has focused on new computational strategies for efficient estimation in constrained latent class models (Mooijaart and van der Heijden, 1992; Hoijtink, 1998).

In this article we propose a methodology for fitting latent class models to high-dimensional data. To this end we employ a *penalized latent class model* that constrains the parameters in order to regularize estimation, but does not make the unrealistic as-

sumption that class-specific parameters are constant either within or between classes. We develop the methodology as follows. In Section 2 we review latent class models and pose them in a regression context. In Section 3 we extend the model to accommodate constrained estimation using a penalized likelihood, and in Section 4 we propose an alternative parameterization of conditional probabilities to concentrate support on a smaller number of parameters. In Sections 5 and 6 we discuss fitting and penalty selection, respectively. In Section 7 we report the results of a simulation study. In Section 8 we present an analysis of the LOH data described above using our proposed methodology. We conclude with some closing remarks in Section 9.

2 Latent Class Model

Assume we observe a sample of n subjects $i \in \{1, ..., n\}$, and from each subject i we have collected $m_i \leq m$ dichotomous variables Y_{ij}, $j \in \mathcal{J}_i \subset \{1, ..., m\}$. Since we are interested in pooling subjects from distinct populations, we further assume that each subject belongs to an *observed* group. That is, for each subject i we observe a group membership indicator G_i, which takes values $g = 1, ..., n_G$. In our example, we have two groups of tumors: the MGH tumors and the HFH tumors. However, there also exists an *unobserved* latent class membership indicator K_i, which takes values on $k = 1, ..., \kappa$, conditional on observed group membership:

$$\mathrm{P}(K_i = k | G_i = g) = \eta_{kg}. \tag{1}$$

The number of classes κ is unknown, although in practice it is treated as fixed, as we discuss in detail in Section 6. In (1), each η_{kg} is unknown and one of our objectives is to estimate the collection of these probabilities. Only $n_G(\kappa - 1)$ parameters are non-redundant, since for each g, $1 = \sum_{k=1}^{\kappa} \eta_{kg}$. Assuming $\eta_{\kappa g}$ is obtained as $\eta_{\kappa g} = 1 -$

$\sum_{k=1}^{\kappa-1} \eta_{kg}$, we denote the collection of parameters as $\eta = (\eta_{11}, ..., \eta_{(\kappa-1)n_G})'$.

In our example, there are $n = 93$ brain tumor subjects with up to 19 LOH values each. Each subject had missing LOH variables due to homozygosity. Since homozygosity is presumably independent of tumorigenesis, we view such missing values as missing completely at random (MCAR). Class k refers to an unobserved LOH "profile" that exists in a subset of patients. We have two groups: the "MGH" group from the London Regional Cancer Centre and Massachusetts General Hospital and the "HFH" group from Henry Ford Hospital. It has been assumed that there are two classes of patients based on assessment of LOH at three distal markers: those with loss of the entire chromosome 1p and those without loss of chromosome 1p (Cairncross *et al.*, 1998). One question of interest based on assessment of LOH at several markers along the entire chromosome arm is whether there is heterogeneity across the chromosome and there are actually three or four classes.

Conditional on class membership, the m_i variables for subject i are assumed independent and are characterized as

$$\mathrm{P}(Y_{ij} = 1|K_i = k, G_i = g) = \mathrm{P}(Y_{ij} = 1|K_i = k) = p_{jk}, \tag{2}$$

where p_{jk} is one of $m \times k$ conditional LOH probability parameters, one for each marker j and class k. Equation (2) expresses the typical latent class model. However, in the penalized estimation setting, it will be convenient to generalize it by expressing p_{jk} as a function of covariates:

$$\mathrm{P}(Y_{ij} = 1|K_i = k, X_{ij} = x_{ij}) = h(x'_{ij}\beta_k), \tag{3}$$

where x_{ij} is a q-dimensional vector of known covariates and $h(\cdot)$ is a known function, for example the inverse logit function. The β_k are unknown parameters, and we denote their collection as $\beta = (\beta'_1, ..., \beta'_\kappa)'$. Equation (2) is the special case of (3) obtained by

setting the $m \times 1$ vector x_{ij} equal to one of m distinct canonical unit vectors and letting $\beta_{kj} = \text{logit}(p_{jk})$. In Section 4 we will make use of the more general formulation. Thus for each subject we observe $\mathcal{D}_i = \{G_i, (x_{ij}, Y_{ij})_{j=1,\ldots,m_i}\}$ and we are interested in estimating $\theta = (\eta', \beta')'$. We assume that Y_{ij} are independent conditional on class membership.

As in other latent class settings, we use maximum likelihood to obtain estimates. To stabilize numerical optimization, it may be convenient to parameterize $\eta_{\kappa g}$ in such a way that it is constrained to lie within the unit interval: for example $\eta_{kg} = \exp(\eta^*_{kg}) / \sum_{k=1}^{\kappa} \exp(\eta^*_{kg})$, where $\eta^*_{\kappa g} = 0$. For simplicity of exposition, we accommodate this technicality by writing as $\dot{\eta}_{kg}$ the derivative of the constrained parameter η_{kg}, but otherwise ignore the details of the unconstrained parameterization.

We write the log-likelihood function as $L(\eta, \beta) = \sum_{i=1}^{n} L_i(\eta, \beta)$, where

$$L_i(\eta, \beta) = \log \left[\sum_{k=1}^{\kappa} \eta_{kG_i} \prod_{j \in \mathcal{J}_i} \{h(x'_{ij}\beta_k)\}^{Y_{ij}} \{1 - h(x'_{ij}\beta_k)\}^{1-Y_{ij}} \right].$$

The corresponding score functions are

$$\frac{\partial L_i}{\partial \eta_{kg}} = \exp\{-L_i(\eta, \beta)\} 1(G_i = g) \dot{\eta}_{kg} \prod_{j \in \mathcal{J}_i} \{h(x'_{ij}\beta_k)\}^{Y_{ij}} \{1 - h(x'_{ij}\beta_k)\}^{1-Y_{ij}}$$

and

$$\frac{\partial L_i}{\partial \beta_k} = \sum_{k=1}^{\kappa} \pi_{ik}(\eta, \beta) \sum_{j \in \mathcal{J}_i} \left\{ Y_{ij} \frac{\dot{h}(x'_{ij}\beta_k)}{h(x'_{ij}\beta_k)} - (1 - Y_{ij}) \frac{\dot{h}(x'_{ij}\beta_k)}{1 - h(x'_{ij}\beta_k)} \right\} x_{ij},$$

where $1(\cdot)$ is the binary indicator function, $\dot{h}(\cdot)$ is the derivative of $h(\cdot)$, and

$$\pi_{ik}(\eta, \beta) = \frac{\eta_{kG_i} \prod_{j \in \mathcal{J}_i} \{h(x'_{ij}\beta_k)\}^{Y_{ij}} \{1 - h(x'_{ij}\beta_k)\}^{1-Y_{ij}}}{\sum_{k=1}^{\kappa} \eta_{kG_i} \prod_{j \in \mathcal{J}_i} \{h(x'_{ij}\beta_k)\}^{Y_{ij}} \{1 - h(x'_{ij}\beta_k)\}^{1-Y_{ij}}} \tag{4}$$

is the *posterior probability* of membership in class k. Note that the dependence on group is expressed only through the probability η_{kG_i} of class membership conditional on G_i. However, group-specific conditional probabilities can be incorporated by including group indicators in the covariates x_{ij}. Note also that under MCAR assumption, missing variables are easily accommodated, since the variables are independent conditional on class

membership. Once we have obtained estimates $\widehat{\eta}$ and $\widehat{\beta}$, estimated posterior probabilities of class membership are obtained from (4) as $\widehat{\pi}_{ik} = \pi_{ik}(\widehat{\eta}, \widehat{\beta})$.

Certain model comparisons can adequately be achieved by using the likelihood ratio test. Specifically, the likelihood ratio statistic, constructed in the usual manner, can be used to test constraints among the estimated parameters θ when the null space lies in the interior of the space of unrestricted parameters. Alternatively, one might consider an approximate cross-validation procedure, such as the one we will describe in Section 6. The latter approach has wider applicability than the likelihood ratio test, since it can be used when there is no natural nesting to candidate models, and can also be used to compare models with differing values of κ.

3 Penalized Latent Class Model

The methodology described in the previous section is sufficient when q is much smaller than n. Although we observe $\sum_{i=1}^{n} m_i$ variables, the m_i variables corresponding to subject i are independent only conditionally on the unobserved class indicator K_i, so in fact there are only n independent units of observation. Thus, the methodology breaks down when q is larger than n, or even when it is a substantial fraction of n. However, such situations are often of interest, as in the LOH application described above.

We address this dimensionality problem by considering a constrained version of the methodology described above. Consider the following penalized likelihood:

$$L_C(\eta, \beta; \Lambda) = L(\eta, \beta) - C(\beta, \Lambda), \tag{5}$$

where $C(\beta, \Lambda)$ is a nonnegative penalty function dependent upon a $q \times q$ matrix Λ. Examples of reasonable penalty functions are the ridge and lasso penalties (Hastie *et al.*, 2001). The ridge penalty is defined generally as $C(\beta, \Lambda) = \beta' \Lambda \beta$ and is equivalent in form

to $C_2(\beta, \Lambda) = \sum_{k=1}^{\kappa} \sum_{j=1}^{q} \lambda_{kj} |\beta_{kj}|^2$ when Λ is diagonal. The alternative lasso penalty takes the form $C_1(\beta, \Lambda) = \sum_{k=1}^{\kappa} \sum_{j=1}^{q} \lambda_{kj} |\beta_{kj}|$. More generally, when Λ is diagonal it is possible to consider a family of L_p penalties of the form $C_p(\beta, \Lambda) = \sum_{k=1}^{\kappa} \sum_{j=1}^{q} \lambda_{kj} |\beta_{kj}|^p$, where $p \geq 1$. The ridge penalty is useful in situations where $x'_{ij}\beta$ varies "smoothly", while the lasso penalty is more useful in "sparse" situations where most elements of β are small in magnitude (Tibshirani, 1996).

In the constrained setting, the score function for β has an additional term,

$$\frac{\partial L_C}{\partial \beta} = \frac{\partial L}{\partial \beta} - \frac{\partial C}{\partial \beta}.$$

For the ridge penalty, $\partial C / \partial \beta = 2\Lambda\beta$. For the lasso penalty, $\partial C / \partial \beta$ is a piecewise constant function, with singularities where $|\beta_{kj}| = 0$; this presents computational challenges, as we describe below.

A priori, it is impossible to know what the value of the constraint parameter Λ should be to achieve optimal results. However, in anlaogy with Hastie and Tibshirani (1990), the use of prediction error obtained from an approximate cross-validation procedure can inform the choice of Λ. Alternatively, as we discuss in Section 6, computationally efficient criteria such as the Akaike Information Criterion (AIC, Akaike, 1974) or the Bayesian Information Criterion (BIC, Schwartz, 1978) may be used.

4 Orthogonal Transformations

It is desirable to leave an "intercept" unconstrained for every class k, thus allowing the classes to distinguish themselves at least in mean response. This is difficult to do using the simple parameterization represented by the special case of (3) wherein the vectors x_{ij} are unit indicators. In addition, it is often of interest to embed a model that contrasts important features of the data within a larger model that captures finer details. However,

the contrasts may involve a large number of the dimensions of the covariate vector x_{ij}. If the support of β_k is distributed among most or all q dimensions of these dimensions, then a large diagonal penalty matrix Λ may impose severe bias on the estimates $\widehat{\beta}_k$, since it could lead to shrinkage of every coefficient. However, if the conditional probabilities (3) could be reparameterized so that the support of β_k is concentrated on a small number of dimensions, the bias imposed upon $\widehat{\beta}_k$ would be less severe, especially if the number of nonzero parameters are small enough to leave unpenalized. To this end, we apply orthogonal transformations to the linear model appearing in (3).

Consider a $q \times q$ matrix U such that $U'U = I$, and an alternative parameterization

$$\mathrm{P}(Y_{ij}|K_i = 1, X_{ij} = x_{ij}) = h(x'_{ij}\beta_k) = h(x'_{ij}U'U\beta_k) = h(x'_{ij}U'\beta^*_k),$$

where $\beta^*_k = U\beta_k$. If U induces a dimension reduction in the sense that the coordinates of β^*_k are small for all but a few dimensions, then the method described in Section 3 will tend to estimate β_k with less bias and prediction error. A useful application of this fact involves choosing contrasts between features of interest in the data set. This allows the support of β_k to be distributed among a small number of coordinates that correspond to *features* of direct interest to the investigator. Assuming these contrasts can be made orthogonal to one another, a full set of orthogonal contrasts can be obtained by augmenting the feature constrasts with *detail* vectors obtained through Gram-Schmidt orthogonalization.

To make ideas concrete, consider the example described in the Introduction. Of nineteen genetic markers, fifteen lie on chromosome 1p and four lie on chromosome 19q. The first five markers represent the distal tip of 1p, where the three markers traditionally used for analysis reside, while the next ten represent locations distributed along the remaining portion of the chromosome arm. Thus, contrasts involving these chromosomal locations are of interest.

In the brain tumor application, the mean probability of LOH can be extracted from

the unit-length vector $u_1 = 19^{-1/2} J_{19}$, where $J_d \in \mathbb{R}^d$ denotes a vector of ones. A contrast comparing probabilities between the two chromosomes is obtained as $u_2 = \omega_2(4J'_{15}, -15J'_4)'$, which is orthogonal to u_1 and can be scaled to have unit length with an appropriate choice of ω_2. A contrast comparing probabilities between the distal tip and central locations of chromosome 1 is obtained as $u_3 = \omega_3(10J'_5, -5J'_{10}, O'_4)'$, where $O_d \in \mathbb{R}^d$ is a vector of zeros; and $u_4 = \omega_4(O'_5, 5J'_5, -5J'_5, O'_4)'$ contrasts probabilities between markers 5-10 and 11-15. Assuming each ω_j is chosen appropriately, the set $\{u_1, u_2, u_3, u_4\}$ comprise an orthonormal set of contrast vectors in $\mathbb{R}^{19}$. It is straightforward to augment these four vectors with fifteen others that are easily interpretable as contrasts of finer details among the 19 markers and which, together with the four vectors just described, comprise an orthonormal set of vectors. Thus the matrix $U = (u_1, u_2, ..., u_4)'$ is an orthonormal feature matrix that concentrates distinctions between different domains of the chromosome in a handful of coordinates. If the investigators are primarily interested in looking for constrasts among these features, constrained analysis using the parameterization $x'_{ij} U' \beta^*_k$ will tend to provide more satisfactory results in terms of bias and prediction error, as we demonstrate in Section 7. In larger problems, it may not be practical to work out the detail contrasts by hand; in such cases, the Gram-Schmidt orthogonalization algorithm is useful.

We remark that this feature-based parameterization facilitates the imposition of appropriate constraints on the model. For example, let Λ_1 be the diagonal matrix with ones corresponding to feature contrasts u_2, u_3, and u_4 in each class, and zeros everywhere else, and let Λ_2 be the diagonal matrix with ones corresponding to the detail contrasts u_5 through u_{19} and zeros everywhere else. If $\alpha_1 > \alpha_2$, then the penalty matrix $\Lambda = \alpha_1 \Lambda_1 + \alpha_2 \Lambda_2$ constrains the detail contrasts more than the feature contrasts.

The proposed transformation is similar in spirit to the *discrete wavelet transform* pro-

posed by Morris *et al.* (2003). Wavelets, originating from image processing theory, are an anologue to Fast Fourier Transforms adapted to data characterized by numerous spikes. Morris *et al.* (2003) imposed penalties on different levels of the wavelet decomposition, much in the same way that in our context (α_1, α_2) imposes different penalties on features and details.

5 Parameter Estimation and Inference

When $C(\beta, \Lambda)$ is differentiable, (5) can be maximized either directly using a quasi-Newton method or using a variant of the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977) described by Bandeen-Roche *et al.* (1997) for latent class problems. In the latter algorithm, estimates of the posterior probabilities (4) are obtained in the expectation step and are used in the maximization step to obtain a provisional solution $\widehat{\beta}$ and $\widehat{\eta}$. For $p = 2$ (ridge penalties), the conditional probability parameter $\widehat{\beta}$ can quickly be obtained using iteratively-reweighted least squares (IRLS). For $p = 1$ (lasso penalties), the penalized likelihood is not differentiable. Consequently, following the recommendation of Tibshirani (1996), we use quadratic programming to successively minimize approximations to the penalized likelihood. We will refer to this algorithm as iteratively-reweighted quadratic programming (IRQP). The EM, IRLS, and IRQP algorithms are sketched in the Appendix.

As described in both Goodman (1974) and Bandeen-Roche *et al.* (1997), the model proposed in Section 2 is not globally identifiable. Bandeen-Roche *et al.* (1997) proposed conditions under which a similar model is *locally* identifiable, i.e. identifiable in a neighborhood of a parameter value. A practical consequence is that the computed maximum of (5) can be sensitive to the initial values supplied to the optimization algorithm. Many commercial software packages use multiple, randomly-generated initial values to increase

the probability of finding a global maximum. We also recommend starting from several different initial values, although the computational demands described in Section 6 limit the number of initial values that can be considered for every candidate Λ. Instead, we focus on choosing a handful of initial values carefully, and sketch reasonable approaches in the Appendix.

Conditional on the penalty Λ and the number of classes κ, it is possible to compute standard errors for the parameters η and β. By Taylor expansion, an estimate of the variance of the estimator $(\eta', \beta')'$ obtained by fixing Λ is simply $H_C^{-1} n V H_C^{-1}$, where H_C is the Hessian matrix of L_C and V is an estimate of the asymptotic variance of the score component $(\partial L_i/\partial\eta, \partial L_i/\partial\beta)$. For the ridge penalty, $H_C = H + 2\Lambda$. In low dimensions, nV is well-approximated by H, the Hessian matrix of L evaluated at $\widehat{\theta}$. In higher dimensions or smaller sample sizes, it may be preferable to estimate nV empirically from the data. As in the estimating equations literature, we refer to standard errors obtained by latter type of estimate as "robust". These standard errors can be used to construct Wald tests in the usual manner.

6 Selection of the Penalty

Estimates of cross-validated prediction error and their computationally efficient approximations are often used for model selection (Hastie *et al.*, 2001, Ch. 7). A computationally efficient approximation is available for comparing distinct models when the constraint function has the form of the L_p family described in Section 3 and $p > 1$. Consider the n-fold cross-validated log-likelihood loss $R = -2\sum_{i=1}^{n} L_i(\widehat{\eta}_{-i}, \widehat{\beta}_{-i})$, where $\widehat{\theta}_{-i} = (\widehat{\eta}'_{-i}, \widehat{\beta}'_{-i})'$ has been obtained, using the same model, by deleting the i^{th} observation. This is the AIC adapted to the present setting. It is straightforward to demonstrate that

R is approximated by

$$\widehat{R} = 2\mathrm{tr}(H_C^{-1} nV) - 2L(\eta, \beta), \tag{6}$$

where, as in section 5, H_C is the Hessian matrix of L_C and V is an estimate of the asymptotic variance of the score component.

In low dimensions, nV is well-approximated by H. This leads to a classical AIC, in which $\mathrm{tr}(H_C^{-1} H)$ represents an effective degrees-of-freedom quantity, identical in form to that given by Ruppert *et al.* (2003, Chaper 8.3) for additive models, and is equal to the number of parameters when the likelihood is unconstrained. However, the approximation may break down in high dimensions, so it may be preferable to estimate nV empirically from the data. We remark that using the empirical estimate of nV in finite samples, unconstrained likelihoods can yield degrees-of-freedom not necessarily equal to the number of parameters.

When n is relatively small, the AIC can be a somewhat poor approximation to R. In particular, it can lead to overfitting (Hurvich and Tsai, 1995). Therefore, it may be preferable to use an alternative such as $\mathrm{BIC} = d\log(n) - 2L(\eta, \beta)$, where d is a degrees-of-freedom quantity that we compute as $\mathrm{tr}(H_C^{-1} nV)$. In Section 7 we offer recommendations on the best choice among AIC, BIC, and three other types of *information criterion* (IC): the bias-corrected AIC, $\mathrm{AIC(c)} = 2d + 2d(d+1)/\max(0, n-d-1) - 2L(\eta, \beta)$ (Sugiura, 1978); $\mathrm{HQIC} = 2d\log[\log(n)] - 2L(\eta, \beta)$ (Hannan and Quinn, 1979); and $\mathrm{CAIC} = d\log(n+1) - 2L(\eta, \beta)$ (Bozdogan, 1987). Generalized Cross Validation (GCV), which uses a mean-squared-error (MSE) loss function, is another typical choice employed for penalization problems. However, difficulties in constructing a computationally simple approximation to the MSE in this context motivate the use of the log-likelihood loss function and related, easily computed quantities.

It is usually sufficient to consider a small number r of known diagonal matrices

$\Lambda_1, ..., \Lambda_r$, set $\Lambda = \sum_{s=1} \alpha_s \Lambda_s$, and search for an optimal $\alpha' = (\alpha_1, ..., \alpha_r)$. Note that setting $\alpha = 0$ corresponds to the absence of constraint, with the constraint becoming more severe as the elements of α grow large. In most applications, the smallest reasonable model involves a nonzero predictor. Consequently, an intercept term in β_k should usually be left unconstrained. Thus, reasonable choices for Λ_s are usually restricted to diagonal matrices with entries equal to zero for the coordinates corresponding to each of the κ intercepts, and positive for some of the other coordinates.

The penalized likelihood is not differentiable for the L_1 penalty, so approximation (6) fails. However Tibshirani (1997) and Wahba (1980) describe a useful approximation, obtained by setting $H_C = H + W$, where W is zero everywhere except for the diagonal elements corresponding to nonzero coefficients β_{kj}, which are set to $|\beta_{kj}|^{-1}$. An alternative is to use the limit of H_C as $p \downarrow 1$. In this approximation, W is replaced by the diagonal matrix having nonzero elements equal to ε^{-1} only for the parameters corresponding to $|\beta_{kj}| < \varepsilon$. The latter, which we found gives more sensible degrees-of-freedom, essentially counts the number of non-negligible parameters. The Tibshirani-Wahba approximation tends to produce smaller degrees-of-freedom.

To optimize an IC over various penalty choices, we recommend a grid-search over a coarse grid of candidate penalties. Using the candidate producing the lowest criterion as an initial value, a more refined search can be conducted using, for example, a simplex algorithm. We have found the Nelder-Mead simplex method (Nelder and Mead, 1965) to be adequate for this purpose. A quasi-Newton algorithm is more difficult to implement, since the second derivatives of (6) and related expressions are analytically intractible, and in the case of the lasso penalty do not even exist. In light of cautionary statements against placing too much trust in automatic tuning parameter selection (e.g. Ruppert *et al.*, 2003, Ch. 5.4 and Ch. 8.4), there is little value in refining the precision of α

beyond two decimal places, and the sensibility of the final solution should be validated visually.

Finally, we remark that in practice the number of classes κ is often not known in advance. We recommend an approach similar to profile-likelihood, where the chosen IC is minimized for each feasible value of κ and κ is chosen as the value that produces the minimum criterion overall. Note that although κ could be considered a model parameter, it is difficult to pose its estimation in a likelihood-based setting since the dimension of θ, $n_G(\kappa - 1) + \kappa q$, depends on κ. Our recommendation for choosing κ is similar to existing methods where a goodness-of-fit test is used to compare different κ values (e.g. Bandeen-Roche *et al.*, 1997).

7 Simulations

To study the behavior of our proposed methodology, we conducted several simulations. We considered two separate cases, each illustrated in Figure 1. Each case involved nineteen markers and three underlying classes. In the first case (Case I), one class was uniformly high, and for the remaining two classes the probability of an event varied "smoothly" over the first fifteen markers, then abruptly jumped to a constant value for the remaining four markers; for one class the probabilities were near zero for the middle markers. In the second case (Case II), each class had low event probability everywhere except in one region unique to the class. The first case represents a situation where classes are characterized primarily by overall response level, and the variation in response probabilities across markers is relatively "smooth". The second case represents a situation where classes are characterized primarily by marker-specific responses, and the response probabilities contain big "jumps". In both cases, $\eta' = (0.40, 0.35, 0.25)$ and $h(\cdot)$ was the inverse of the logit function. For each case, we considered 250 simulated data sets, and

for each simulated data set, we set $n = 50$ subjects. There were no missing values in our simulation study.

7.1 Description of Analyses

For each simulated data set in Case I, we conducted a total of fifty-four analyses: for each of $\kappa \in \{2, 3, 4\}$, we conducted six analyses using a naive marker-based parameterization, with an intercept representing the first marker and indicators for the remaining eighteen markers, with ridge penalties $\lambda \in \{0.01, 0.1, 1, 10, 100, 1000\}$, applied to all non-intercept β coefficients; eighteen corresponding analyses using the feature parameterization described in Section 4, with the same κ values and penalties; and eighteen corresponding analyses with the feature parameterization, where the feature contrasts u_2, u_3, and u_4 were penalized using λ while the detail contrasts were penalized using 2λ. We refer to each set of eighteen analyses respectively as the Markers, Features I, and Features II analyses. Note that for the latter two sets, $\Lambda = \alpha_1\Lambda_1 + \alpha_2\Lambda_2$, as defined in Section 4. Thus, in the notation of Sections 4 and 6, the Markers parameterization used a single penalty parameter $\alpha_1 = \lambda$, the Features I parameterization used a bivariate penalty $(\alpha_1, \alpha_2) = (\lambda, \lambda)$, and the Features II parameterization used a bivariate penalty $(\alpha_1, \alpha_2) = (\lambda, 2\lambda)$. For each analysis we computed five types of IC: AIC, BIC, AIC(c), HQIC, and CAIC, as defined in Section 6. For each IC and for each value of κ and parameterization (Markers, Features I, and Features II), we recorded the penalty producing the lowest criterion value. Analysis of the Case II simulations was identical to Case I except that $\lambda \in \{0.001, 0.01, 0.1, 1, 10, 100\}$, since the Case II data sets tended to require smaller penalties. Thus for each of the two cases, there were 3 κ values $\times$ 3 parameterizations $\times$ 6 λ values = 54 analyses.

7.2 Results for All κ Values

For each data set and IC, we tabulated which value of κ produced the lowest value of the IC. Table 1 summarizes the results. In general, AIC(c), BIC, and CAIC were minimized most often by the correct $\kappa = 3$, while AIC and HQIC were more often minimized by the incorrect $\kappa = 4$. Thus, in our simulations, using AIC as a criterion for selection of κ would often lead to overfitting. For each IC and parameterization, Table 1 also presents the number of data sets for which the smallest tuning parameter λ minimized the IC, thus giving an impression of the frequency of data sets for which little penalization was required. For a substantial fraction of the data sets, AIC and HQIC were minimized at the lowest value of λ. All five of the IC types tended to be minimized at smaller values of λ for the Marker parameterization than for the Features I and Features II parameterizations.

7.3 Results for $\kappa = 3$

Fixing $\kappa = 3$, for each IC we selected the value of λ that minimized the IC and compared the corresponding fit to the true parameter values. For the Features I analyses applied to both cases, Figure 1 depicts the mean of the estimated probabilities (from the model obtained from the value of λ that minimized AIC) over each marker, conditional on class, compared with the true values that generated the data. The picture is similar for all three parameterizations. Thus, on average, constrained estimation seems to perform well with respect to estimating conditional probabilities, regardless of the parameterization used.

In order to compare estimated parameters with the true parameters used to generate the data, it was necessary to overcome the familiar "labeling problem" common in latent class analysis (Stephens, 2000). Because there is no natural ordering to the classes, other than that induced by the estimated parameter η, estimated classes had to be matched to

the generating classes in a sensible fashion. To achieve this matching, we minimized the conditional error, $CE = \sum_k \sum_j \{h(x_j'\widehat{\beta}_{\rho(k)j}) - h(x_j'\beta_{kj})\}^2$, with respect to permutations $\rho(\cdot)$ of $\{1, ..., \kappa\}$. This operation, which is similar to a decision-theoretic procedure described by Stephens (2000), insured that a solution was not unfairly penalized for simply mislabelling the classes.

After relabelling the classes, we applied two criteria to evaluate the performance of our methods. These criteria are defined as follows:

$$\text{MSE} = \{\textstyle\sum_k (\widehat{\eta}_k - \eta_k)^2 + \sum_k \sum_j (\widehat{p}_{kj} - p_{kj})^2\} = \textit{mean squared error}, \text{ and}$$

$$\text{LL} = -\textstyle\sum_k \sum_j \eta_k \{p_{kj} \log(\widehat{p}_{kj}) + (1 - p_{kj}) \log(1 - \widehat{p}_{kj})\} = \textit{likelihood loss},$$

where $p_{kj} = x_j'\beta_k$ and $\widehat{p}_{kj} = x_{kj}'\widehat{\beta}_k$. Note that LL, related by a constant to the Kullback-Leibler information assuming that class membership is actually observed, is an analogue to the deviance statistic, and is minimized when $p_{kj} = \widehat{p}_{kj}$ for all j and k.

Table 2 presents a summary of evaluation criteria for the analyses performed with $\kappa = 3$. For each parameterization and evaluation criterion, Table 2(a) shows the number of data sets for which the criterion was minimized by the parameterization. For both cases, for a substantial majority of the data sets, the Features II parameterization minimized both criteria. For each parameterization and evaluation criterion, Table 2(b) gives the number of data sets for which the smallest tuning parameter λ minimized the evaluation criterion. For Case I using the Markers parameterization, MSE was minimized at the lowest value of λ for a majority of the data sets. For the other parameterizations and for Case II, MSE was rarely minimized at the lowest value of λ. In all situations, LL was rarely minimized at the lowest value of λ.

For each parameterization and IC, Table 2(c) shows the mean MSE and LL where, for each parameterization, λ was selected as the minimizer of the IC. Thus, Table 2(c) summarizes the effect of using IC to select the tuning parameter λ. In general, the

Features II parameterization tended to produce smaller values of the evaluation criteria than did the Features I parameterization, which tended to produce smaller values than the Markers parameterization. For each case, IC, and evaluation criterion, the paired t-test comparing the Features II solution to the Markers solution produced a P value less than 0.001, strongly suggesting the superiority of the Features II parameterization. No IC unequivocally produced the smallest values of the evaluation criteria, although AIC(c) tended to produce notably higher values than the other four ICs, especially for the Features II parameterization.

Thus, the Features I and Features II parameterizations seem superior to the Marker parameterization. Note that the two different types of feature-based parameterizations investigated in this simulation study, $\alpha = (\lambda, \lambda)$ and $\alpha = (\lambda, 2\lambda)$, were meant to represent distinct feature-based penalization strategies within the computational constraints imposed by analyzing a large number of data sets. In practice, we recommend searching for the multivariate penalty α that minimizes IC. We demonstrate this in the next section.

We remark that, with the exception of MSE in Case I using the Markers parameterization, the smallest penalty we considered ($\lambda = 0.01$ for Case I and $\lambda = 0.001$ for Case II) failed to minimize the evaluation criteria in over 85% of the simulated data sets. In many cases, the smallest penalty $\lambda = 0.01$ minimized the criterion for a negligible fraction of the simulated data sets. Because the unconstrained model corresponds to the special case $\lambda = 0$, this demonstrates a need for penalized estimation in settings where the sample size is small in comparison with the dimension of the covariate space. The exceptional case was a by-product of the difficulty of using the naive representation (2) in the constrained estimation setting, since the unpenalized coefficient was a "reference" marker that need not be representative of the remaining markers. As Table 2(c) shows, the Markers parameterization produced higher MSE and LL values, demonstrating the

poorer properties of the naive parameterization.

AIC and HQIC were often minimized by $\kappa = 4$ rather than the correct $\kappa = 3$. AIC(c), BIC, and CAIC more consistently minimized at $\kappa = 3$. Thus, AIC(c), BIC, and CAIC were more reliable criteria for selecting the number κ of classes. However, AIC(c) proved less reliable as a criterion for selecting λ. Consequently, we recommend the use of BIC or CAIC as a general criterion for model selection.

8 Application

We applied the methodology described above to the LOH data obtained from MGH and HFH. Using both ridge and lasso penalties, we found the optimal penalties by minimizing BIC for seven different models: $\kappa = 1$, which is simply penalized logistic regression; $\kappa \in \{2, 3, 4\}$ assuming class prevalence η is uniform over the MGH and HFH populations; and $\kappa \in \{2, 3, 4\}$ assuming class prevalence η differs between the MGH and HFH populations. In all cases we used the feature-based parameterization of LOH marker, $h(t) = \text{logit}^{-1}(t)$. and a bivariate penalty parameter (α_1, α_2), with α_1 constraining feature contrasts u_1, u_2, and u_3, as described in Section 4 and α_2 constraining the detail constrasts u_4 through u_{19}. We were unable to fit the model for $\kappa = 4$ when η was assumed to depend on the MGH/HFH grouping. We used four starting values for every value of (α_1, α_2) used in the coarse grid search described in Section 6.

Table 3 displays the results. Note that the "naive" BIC computed with nV approximated by H was typically, but not always, higher than the robust BIC computed with an empirical estimate of nV. Note also that the optimal penalties for $\kappa = 1$ were essentially infinite, indicating that an intercept-only model was better than any other. This was consistent with the results of a likelihood ratio test comparing the unpenalized logistic regression estimates for the full model ($q = 19$) and the intercept-only model ($q = 1$); the

test statistic was 22.46 on 18 d.f., with a P-value of 0.21. The BIC values for $\kappa = 2$, 3 and 4 were much smaller than for $\kappa = 1$. The smallest BIC was produced by $\kappa = 3$ with uniform class prevalence for the MGH and HFH populations. The next smallest BIC value was produced by $\kappa = 3$ with distinct class prevalences for the MGH and HFH subpopulations. From the ridge penalty fit of the latter model, using $\widehat{\Sigma} = nH_C^{-1}\widehat{V}H_C^{-1}$ as an estimate of $\mathrm{Cov}(\widehat{\theta})$ and $\widehat{V}$ as the empirical variance of the score functions, we constructed a Wald test for the equivalence of the prevalences of the two classes. Specifically, the null hypothesis was that $\eta_{k,MGH} = \eta_{k,HFH}$ for $k = 1, 2$. The resulting chi-square value was 7.04 on 2 d.f., yielding a P-value of 0.03. Thus, although it seems clear that $\kappa = 3$ produced the best fit, there was some equivocation in whether the class prevalences differred between MGH and HFH. For the latter model, the MGH Class 1, 2, and 3 prevalences were 0.12, 0.33, and 0.55, respectively, while the corresponding HFH prevalences were 0.25, 0.48, and 0.27. Because the Wald tests suggested heterogeneity in class prevalence with respect to the groups, we chose the latter model, in which the heterogeneity is made explicit, as the best-fitting model.

Figure 2 displays the fitted marker probabilities $h(U'\widehat{\beta}_k)$ from this model, along with robust confidence limits, calculated using the square roots of the diagonal entries of $U'\widehat{\Sigma}U$. Class prevalences are given in the footer of the figure. The figure shows one latent class with uniformly high probability of LOH, and two classes with typically low LOH probabilities. The two low-LOH classes are distinguished by probabilities for 19q and markers closer to the centromere of chromosome 1. Class 1 had probabilities that were practically zero for all but the distal tip of 1p. The corresponding illustration for the model having different prevalences for the MGH and HFH groups is virtually identical with Figure 2. Note that for the latter model, the MGH group tended to have a higher prevalence of Class 3 subjects (0.55) compared with the HFH group (0.27); thus the

MGH population had considerably higher prevalence of subjects with high LOH. The MGH prevalence is similar to the prevalence 0.56 Dong *et al.* (2004) reported for tumors with LOH at all informative markers.

A natural question that arises is whether the classes obtained from our methodology have utility in predicting survival. To answer this question, we conducted an exploration of survival, conditional on class, by fitting a weighted Cox proportional hazards model. In this analysis, we included three rows for each subject, one for each class, and weighted each row with the corresponding posterior probability that the subject was a member of the class. Additionally, we stratified by MGH/HFH population. Table 4 presents the results. Compared with the low LOH Class 1, the high LOH Class 3 had a much better prognosis, with hazard ratio equal to 0.17, 95% confidence limits $(0.06, 0.47)$. In terms of survival, Class 2 was more similar to Class 1; its hazard ratio was insignificant compared with Class 1 but significantly higher than 1.0 when compared with Class 3. These results confirm the survival benefit of uniformly high LOH for oligodendroglioma patents (Cairncross *et al.*, 1998).

9 Closing Remarks

In this article we have proposed a latent class model where, conditional on unobserved membership in one of κ classes, the variables are independent with probabilities determined by a regression model of high dimension q. We address the dimensionality problem by using a family of penalties including the ridge and lasso. Finally, we present an orthogonal map that parameterizes the conditional probabilities as contrasts involving different levels of detail.

Our methodology generalizes the parameter constraints proposed by Lazarsfeld and Henry (1968), Agresti and Lang (1993), and Lindsay *et al.* (1991) in the sense that the

prior methods essentially impose either a zero or an infinite penalty on the conditional probabilities transformed as $h(U'\beta)$ for some appropriate matrix U. Figure 3 depicts the results of a 3-class model fit to the LOH data with a lasso penalty, thus shrinking a majority of the coefficients completely to zero. As is evident from the figure, the resulting fit produced identical probabilities for markers in the same chromosome region, according to the transformation described in Section 4.

For simplicity, we have assumed that the class prevalences η depend at most on a single categorical variable G_i. Aside from computational complexity and the limits imposed by finite sample sizes, there is nothing that prevents the construction of a full regression model on η in the manner proposed by Bandeen-Roche *et al.* (1997). For example, if G_i were a vector of covariates rather than a categorical indicator, the prevalences might be parameterized as $\mathrm{P}(K_i = k|G_i = g) = \tilde{h}(g'\gamma)$. In another application, not reported here, we encountered difficulties in fitting models with a larger number of levels of G_i; consequently, it may be difficult to fit complicated regression models for a small or moderately sized data set. In principle, one could also constrain the regression model $\tilde{h}(g'\gamma)$, although this complicates the search for the optimal penalty.

One question that arises is whether the "true" number of classes κ can be recovered using latent class methodology. Mixture distributions are typically applied in two situations: approximating non-normal distributions and modeling population heterogeneity. In the latter case, classes are thought to correspond to meaningful subpopulations. If the data are non-normal and stem from a heterogeneous population, then additional classes serve only to better approximate the observed distribution. In a sense, for real world samples, κ is in fact equal to n, since no two subjects have exactly the same probabilities of $Y_{ij} = 1$. A more practical question is how large κ should be to sufficiently capture the correlation observed in the data set. As reported in Section 8, the best-fitting logistic

regression model ($\kappa = 1$) appeared to contain only an intercept. However, the BICs for both the ridge and lasso fits were quite large for $\kappa = 1$. In other words, a single-class model was completely inadequate for this data set, since it failed to account for the striking correlation between LOH variables within subjects. Although $\kappa = 2$ produced much lower BICs, the lowest BICs were obtained from $\kappa = 3$. It is not unlikely that with a much larger data set, BIC would be optimized by a larger value of κ. However, with $n = 93$ subjects there was insufficient power to refine the correlation structure beyond $\kappa = 3$.

Another question that arises is whether a penalty is necessary at all. Conceiveably, we might have applied the unconstrained latent class methodology proposed by Bandeen-Roche *et al.* (1997). However, depending upon the initial value supplied to the algorithm, unpenalized versions of the 3-class models presented in Section 8 either would diverge or would produce Hessian matrices that were not positive definite. Thus, in our application it appears that a penalty was required to obtain sensible results. The simulations presented in Section 7, notably Table 2(b), support the utility of penalized latent class models.

We proposed the use of summary information criteria to select the values of α and κ, and from simulations determined that BIC or CAIC was generally more reliable than AIC or AIC(c). A reviewer remarked that a nonparametric bootstrap would be a more reliable alternative (e.g. von Davier, 1997). While this is undoubtedly true, the computational demands of our method make the use of resampling difficult.

We were unable to fit a model with $\kappa = 4$ and η dependent on group membership, even though we used multiple starting values. The problem seemed to be that the response probability was essentially zero for one class. Since this corresponded to an infinite value of an unpenalized intercept, the algorithm diverged. In the penalized likelihood

setting, this phenomenon is somewhat unique to latent class analysis; in typical penalized regression settings, a fit can usually be found with large enough penalty. One solution might be to penalize the intercepts at yet another level of penalization, leading to a trivariate α parameter. However, given that the BIC value for the simpler $\kappa = 4$ model was higher than the corresponding model with $\kappa = 3$, we decided that the additional complexity was not worth the effort.

Our methodology could be extended in several different directions. As mentioned above, it is straightforward in principle to impose a regression model on the prevalence parameters η, possibly with an additional penalty. A much more challenging problem is to extend the method to account for polytomous variables Y_{ij} or mixed discrete and continuous variables. The models could also be extended to allow for conditional dependence within classes, along the lines of Qu *et al.* (1996).

Although we used a specific LOH application to motivate and demonstrate our proposed methodology, penalized latent class analysis can be applied in other genomic settings. For example, we are currently applying our methodology to problems involving gene methylation and protein expression. A detailed description of these analyses will appear in separate manuscripts. In the gene methylation example, which focuses on environmental influencs of gene methylation, the covariate-dependent prevalence models popularized by Bandeen-Roche *et al.* (1997) play a central role. Overall, we anticipate that our methodology will be an attractive approach for applied genomic and proteomic problems where expression profiles are of primary interest.

Acknowledgements

This research was supported in part by grants CA075971, CA105956 and ES012044 from the US National Institutes of Health and National Cancer Institute. The authors thank O. Bogler, J.G. Cairncross, and D.N. Louis for use of the LOH data and for

helpful feedback, and two anonymous referees for helpful comments that improved our methodology.

References

Agresti, A. and Lang, J. B. (1993) Quasi-symmetric latent class models, with application to rater agreement. *Biometrics*, **49**, 131–139.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, **AC-19**, 716–723.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L. and Rathouz, P. J. (1997) Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, **92**, 1375–1386.

Bartholomew, D. J. (1987) *Latent Variable Models and Factor Analysis*. New York: Oxford University Press.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **57**, 289–300.

Bozdogan, H. (1987) Model selection and akaike's information criterion (aic): the general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Cairncross, J. G., Ueki, K., Zlatescu, M. C., Lisle, D. K., Finkelstein, D. M., Hammond, R. R., Silver, J. S., Stark, P. C., Macdonald, D. R., Ino, Y., Ramsay, D. A. and Louis, D. N. (1998) Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. *Journal of the National Cancer Institute*, **90**, 1473–1479.

Cawkwell, L., Bell, S. M., Lewis, L. A., Dixon, M. F., Taylor, G. R. and Quirke, P. (1993) Rapid detection of allele loss in colorectal tumours using microsatellites. *Br J Cancer*, **67**, 1262–1267.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Dong, Z., Pang, J. C. S., Ng, M. H., Poon, W. S., Zhou, L. and Ng, H. K. (2004) Identification of two contiguous minimally deleted regions on chromosome 1p36.31 – p36.32 in oligodendroglial tumours. *British Journal of Cancer*, **91**, 1105–1111.

Goldfarb, D. and Idnani, A. (1983) A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, **27**, 1–33.

Goodman, L. A. (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

Hannan, E. J. and Quinn, B. G. (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, **41**, 190–195.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. Boca Raton, Florida: Chapman & Hall/CRC.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

Hoijtink, H. (1998) Constrained latent class analysis using the Gibbs sampler and posterior predictive p-values: Applications to educational testing. *Statistica Sinica*, **8**, 691–711.

Hurvich, C. M. and Tsai, C. L. (1995) Model selection for extended quasi-likelihood models in small samples. *Biometrics*, **51**, 1077–1084.

Ino, Y., Betensky, R. A., Zlatescu, M. C., Sasaki, H., Macdonald, D. R., Stemmer-Rachamimov, A. O., Ramsay, D. A., Cairncross, J. G. and Louis, D. N. (2001) Molec-

ular subtypes of anaplastic oligodendroglioma: Implications for patient management at diagnosis. *Clinical Cancer Research*, **4**, 839–845.

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data – An Introduction to Cluster Analysis*. New York: Wiley.

Lazarsfeld, P. F. and Henry, N. W. (1968) *Latent Structure Analysis*. Boston: Houghton Mifflin.

Lindsay, B., Clogg, C. and Grego, J. (1991) Semiparametric estimation in the rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, **86**, 96–107.

Meulders, M., Boeck, P. D., Kuppens, P. and Mechelen, I. V. (2002) Constrained latent class analysis of three-way three-mode data. *Journal of Classification*, **19**, 277–302.

Mooijaart, A. and van der Heijden, P. G. M. (1992) The EM algorithm for latent class analysis with equality constraints. *Psychometrika*, **57**, 261–269.

Morris, J. S., Vannucci, M., Brown, P. J. and Carroll, R. J. (2003) Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, **98**, 573–583.

Nelder, J. A. and Mead, R. (1965) A simplex algorithm for function minimization. *Computer Journal*, **7**, 308–313.

Qu, Y., Tan, M. and Kutner, M. H. (1996) Random effect models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, **52**, 797–810.

R Development Core Team (2004) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-00-3.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.

Schwartz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Stephens, M. (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B*, **62**, 795–809.

Sugiura, N. (1978) Further analysis of data by Akaike's information criterion and finite corrections. *Communications in Statistics Part A – Theory and Methods*, **7**, 13–26.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistics Society, Series B*, **58**, 267–288.

— (1997) The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385–395.

von Davier, M. (1997) Bootstrapping goodness-of-fit statistics for sparse categorical data: results of a Monte Carlo study. *Methods of Psychological Research*, **Vol. 2**, No. 2.

Wahba (1980) Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. In *International Conference on Approximation theory in honour of George Lorenz*. Austin, Texas: Academic Press.

Westfall, P. H. and Young, S. S. (1993) *Resampling-based multiple testing: examples and methods for p-value adjustment*. New York: Wiley.

A Appendix

In this Appendix we sketch the EM, IRLS and IRQP algorithms described in Section 5, as well as an algorithm for obtaining starting values.

Our EM algorithm is a straightforward generalization of Bandeen-Roche *et al.* (1997):

EM-1. Compute the posteriors $\tilde{\pi}_{ik}$ (4) for each subject i.

EM-2. Within each class k, use IRLS or IRQP to obtain $\tilde{\beta}_k$ weighting each subject by $\tilde{\pi}_{ik}$.

EM-3. Set $\tilde{\eta}_{kg}$ equal to $\sum_i \tilde{\pi}_{ik} 1(G_i = g) / \sum_i 1(G_i = g)$.

EM-4. Iterate steps EM-1 through EM-3 until a convergence criteria is met.

As a criterion for the ridge problem, we used the absolute value of the penalized score functions, that is the derivatives of (5) evaluated at the current iterate. For the lasso problem, we used the absolute value of the change in likelihood between iterations, since the score functions are discontinuous and in finite samples it may not be possible to obtain a score exactly equal to zero.

We use the standard IRLS algorithm adapted for penalized estimation. Assuming $h(\cdot) = \text{logit}^{-1}(\cdot)$ and weights $\tilde{\pi}_{ik}$, the algorithm is as follows.

IRLS-1 Set $\mu_{ij} = h(x'_{ij}\tilde{\beta}_k)$ and $\dot{\mu}_{ij} = \mu_{ij}(1 - \mu_{ij})$.

IRLS-2 Compute $e_{ij} = Y_{ij} - \mu_{ij}$

IRLS-3 Set W equal to $2\Lambda + \sum_i \sum_j \tilde{\pi}_{ik} \dot{\mu}_{ij} x_{ij} x'_{ij}$

IRLS-4 Set $z = \sum_i \sum_j \tilde{\pi}_{ik} e_{ij} x_{ij} - \Lambda \tilde{\beta}_k$.

IRLS-5 Set $\delta = W^{-1} z$ and reset $\tilde{\beta}_k$ equal to $\tilde{\beta}_k + \delta$.

IRLS-6 Iterate Steps IRLS-1 through IRLS-5 until δ is small.

The derivation of this algorithm is a straightforward adaptation of the usual Taylor's expansion used to justify IRLS for generalized linear models. The adaptation for link functions $h(\cdot)$ other than the antilogit function is also straightforward.

The IRQP algorithm replaces step IRLS-5 in IRLS with a quadratic programming minimization step. We use the dual method of Goldfarb and Idnani (1983), found in the *quadprog* library for R. Again assuming $h(\cdot) = \text{logit}^{-1}(\cdot)$, the algorithm is as follows:

IRQP-0 Set $\tilde{\sigma}$ equal to the vector $\text{sign}(\tilde{\beta}_k)$, where the sign function is applied element-wise.

IRQP-1 Set $\mu_{ij} = h(x'_{ij}\tilde{\beta}_k)$ and $\dot{\mu}_{ij} = \mu_{ij}(1 - \mu_{ij})$.

IRQP-2 Compute $e_{ij} = Y_{ij} - \mu_{ij}$

IRQP-3 Set σ equal to the vector of values σ_l, where

$$\sigma_l = \text{sign}(\tilde{\beta}_{kl})1(\beta_{kl}/ne0) - \tilde{\sigma}_l 1(\beta_{kl} = 0).$$

Note that $\sigma = \text{sign}(\tilde{\beta}_k)$ except for indices l for which $\beta_{kl} = 0$; in the latter case, σ_l reverses the sign of $\tilde{\sigma}_l$.

IRQP-4 Set $\tilde{\sigma} = \sigma$.

IRQP-5 Set W equal to $\sum_i \sum_j \tilde{\pi}_{ik}\dot{\mu}_{ij}x_{ij}x'_{ij}$

IRQP-6 Set $z = \sum_i \sum_j \tilde{\pi}_{ik}e_{ij}x_{ij} - \Lambda\sigma$.

IRQP-7 Set $z_0 = -\sigma'\tilde{\beta}_k$

IRQP-8 Set δ equal to the quadratic programming minimum of $\delta^* W \delta^* - z'\delta^*$ subject to the constraint $\text{diag}(\sigma)\delta^* \geq z_0$.

IRQP-9 Reset $\tilde{\beta}_k$ equal to $\tilde{\beta}_k + \delta$.

IRQP-10 Iterate IRQP-1 through IRQP-9 until δ is small or stationary.

Note that, analogous to the IRLS procedure, IRQP-8 obtains a correction to $\tilde{\beta}_k$ by minimizing the second order approximation to the penalized likelihood. The constraint in IRQP-8 prevents $\tilde{\beta}_{kl}$ from switching signs in IRQP-9, shrinking small coefficients to zero. However, in the IRQP-3 step of the next iteration, σ reverses the sign associated with any shrunk coefficient, allowing the coefficient to reverse sign also. Because IRQP is more computationally intensive than IRLS, the algorithm should start with a reasonably accurate estimate (for example, the solution using a ridge penalty).

As mentioned in Section 5, latent class models are quite sensitive to starting values, so many authors have recommended maximizing the likelihood from multiple starting values and selecting the solution that has the maximum likelihood over all solutions. In our experience, starting values should attempt to represent multiple "types" of solutions. For example, a solution may tend to classify subjects by the overall level of response across all markers, or it may tend to classify subjects by high response in particular markers. The two cases simulated in Section 7 represent these two situations.

Starting values for the first case can be obtained using the following algorithm:

Start-I-1 Compute the mean $\overline{Y}_{i\cdot}$ over each subject i and use the results to categorize subjects into κ classes of approximately equal size.

Start-I-2 Set $\tilde{\eta}_{kg}$ equal to the proportion of each class k in group g.

Start-I-3 Set $\tilde{\beta}_{k1}$ (the coefficient corresponding to the intercept) equal to $h^{-1}(\overline{Y}_{i\cdot})$ and $\tilde{\beta}_{kl} = 0$ for $l > 1$.

If q is small enough relative to the sample size, then an alternative to Step Start-I-3 is to set $\tilde{\beta}_k$ equal to its unpenalized IRLS estimate using the initial weights; or if an appropriate transformation as described in Section 4 has been applied, another alternative to Step Start-I-3 is to compute the logistic regression for a smaller-dimensional subset of the covariates representing a coarse parameterization of the covariate-space.

Starting values for the second case can be obtained by first applying a nonparametric classifier to the *markers* (rather than the subjects); then computing $\overline{Y}_{i\cdot}$ within each cluster of markers to obtain class-specific weights for each subject; and finally, using the weights, obtain coarse regression estimates by a weighted logistic regression. A slightly more detailed algorithm is as follows:

Start-II-1 Compute a distance matrix for the markers, for example by subtracting a pairwise correlation matrix from 1.

Start-II-2 Classify the markers using a nonparametric classifier.

Start-II-3 Within each class k obtained from Start-II-2, compute subject-specific weights $\overline{Y}_{i\cdot}^{(k)}$ for each subject.

Start-II-4 For each class k, use the weights obtained in Step Start-II-3 to fit a weighted logistic regression to all covariates, or else a subset of the covariates representing a coarse parameterization of the covariate-space (and setting the remaining covariates to zero).

It is possible to use a "fuzzy" classifier in Step Start-II-2 and a corresponding weighted mean in Start-II-3. For the simulations in Section 7 and the application in Section 8 we used the clustering algorithm adapted from Kaufman and Rousseeuw (1990, Ch. 4) and available in the R software package (R Development Core Team, 2004). However, for the Case I simulations and also for the application, the results typically did not produce starting values that were meaningfully different from the first algorithm. For the Case II simulations, the starting values from the second algorithm tended to produce better results.

In general, we recommend also applying random perturbations to the starting values obtained from the two algorithms described above.

Table 1: Summary of Simulation Results by Information Criterion

(a) Distribution of best κ value as selected by five information criterion values

	Case I					Case II				
κ	AIC	AIC(c)	BIC	HQIC	CAIC	AIC	AIC(c)	BIC	HQIC	CAIC
2	2	15	7	11	7	0	0	0	0	0
3	52	170	173	146	171	17	250	165	69	166
4	196	65	70	93	72	233	0	85	181	84
Total	250	250	250	250	250	250	250	250	250	250

(b) Number of data sets for which smallest value of tuning parameter λ was selected

	Case I					Case II				
Parameterization	AIC	AIC(c)	BIC	HQIC	CAIC	AIC	AIC(c)	BIC	HQIC	CAIC
Markers*	213	0	26	145	24	44	0	9	25	9
Features I	149	0	4	53	4	43	0	0	8	0
Features II	99	0	0	21	0	43	0	0	6	0

Summary of latent class analyses of 250 simulated data sets. Two separate cases were considered, each illustrated in Figure 1. Each case involved 19 markers and 3 underlying classes. For each data set there were $n = 50$ subjects. For each simulated data set a total of 54 analyses were conducted: for each $\kappa \in \{2, 3, 4\}$, 6 analyses using the naive marker-based parameterization with different ridge penalties λ; 18 additional analyses using the feature parameterization described in Section 4, with the same penalties and κ values; and 18 analyses with the feature parameterization, where the feature contrasts u_1, u_2, and u_3 were penalized using the value λ and the remaining detail contrasts were penalized using the value 2λ. Each of the three parameterizations is referred to respectively as Markers, Features I, and Features II. For Case I, $\lambda = 0.01$, 0.1, 1, 10, 100, and 1000. For Case II, $\lambda = 0.001$, 0.01, 0.1, 1, 10, and 100. Table 1(a) gives the distribution of the κ value selected by each of the following five information criteria:

$\text{AIC} = 2d - 2L(\eta, \beta)$ (Akaike, 1974),

$\text{AIC(c)} = 2d + (2d(d+1))/\max(0, n-d-1) - 2L(\eta, \beta)$ (Sugiura, 1978),

$\text{BIC} = d\log(n) - 2L(\eta, \beta)$ (Schwartz, 1978),

$\text{HQIC} = 2d\log[\log(n)] - 2L(\eta, \beta)$ (Hannan and Quinn, 1979),

$\text{CAIC} = d\log(n+1) - 2L(\eta, \beta)$ (Bozdogan, 1987),

where $d = nH_C^{-1}\widehat{V}$ is the "robust" effecitive d.f. defined in Section 6. Table 1(b) gives the number of data sets for which the smallest value of the tuning parameter λ was selected ($\lambda = 0.01$ for Case I and $\lambda = 0.001$ for Case II), thus giving an impression of the number of data sets for which little penalization was required.

*Using the Marker parameterization, 1 Case I data set and 2 Case II data sets could not be fit.

Table 2: Summary of Evaluation Criteria for Simulations

(a) Number of data sets for which parameterization produced minimum criterion value

	Criterion	Markers	Features I	Features II
Case I	MSE	3	10	237
	LL	1	7	242
Case II	MSE	0	2	248
	LL	0	0	250

(b) Number of data sets for which smallest λ produced minimum criterion value

	Criterion	Markers	Features I	Features II
Case I	MSE	159	7	5
	LL	15	0	1
Case II	MSE	26	0	0
	LL	3	0	0

(c) Mean (SE) of criterion value for parameter estimates corresponding to the λ yielding the smallest IC

		Case I					Case II				
		AIC	AIC(c)	BIC	HQIC	CAIC	AIC	AIC(c)	BIC	HQIC	CAIC
	Markers*	0.73	1.09	1.06	0.81	1.06	0.61	1.70	0.64	0.62	0.64
M		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.09)	(0.12)	(0.08)	(0.09)	(0.08)
S	Features I	0.73	1.01	0.87	0.67	0.87	0.33	1.07	0.32	0.32	0.32
E		(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Features II	0.67	0.80	0.67	0.58	0.68	0.29	1.02	0.28	0.28	0.28
		(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
	Markers*	9.29	9.44	9.44	9.32	9.44	7.51	7.67	7.11	7.34	7.11
		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.11)	(0.08)	(0.11)	(0.11)	(0.11)
L	Features I	9.26	9.34	9.22	9.12	9.23	7.02	7.20	6.65	6.71	6.65
L		(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.04)	(0.01)	(0.01)	(0.03)	(0.01)
	Features II	9.09	9.15	9.03	8.96	9.04	6.85	7.17	6.56	6.61	6.56
		(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.04)	(0.01)	(0.01)	(0.02)	(0.01)

Summary of evaluation criteria applied to constrained latent class analyses of 250 simulated data sets. Two separate cases were considered, each illustrated in Figure 1. Each case involved 19 markers and 3 underlying classes. For each data set there were $n = 50$ subjects. For $\kappa = 3$, 18 analyses were conducted: 6 analyses using the naive marker-based parameterization with different ridge penalties λ; 6 additional analyses using the feature parameterization described in Section 4, with the same penalties; and 6 analyses with the feature parameterization, where the feature contrasts were penalized using the value λ and the remaining detail contrasts were penalized using the value 2λ. Each of the three parameterizations is referred to respectively as Markers, Features I, and Features II. For Case I, $\lambda =$ 0.01, 0.1, 1, 10, 100, and 1000. For Case II, $\lambda = 0.001$, 0.01, 0.1, 1, 10, and 100. Table 2(a) gives the number of simulated data sets for which the parameterization represented by the column produced the smallest value of the following two evaluation criteria:

$$\text{MSE} = \text{root mean square error, } \{\textstyle\sum_k(\widehat{\eta}_k - \eta_k)^2 + \sum_k \sum_j (\widehat{p}_{kj} - p_{kj})^2\}$$

$$\text{LL} = \text{likelihood loss, } -\textstyle\sum_k \sum_j \eta_k \{p_{kj} \log(\widehat{p}_{kj}) + (1 - p_{kj}) \log(1 - \widehat{p}_{kj})\},$$

where $p_{kj} = x_j' \beta_k$ and $\widehat{p}_{kj} = x_{kj}' \widehat{\beta}_k$. Note that LL is an analogue to likelihood loss, and is minimized when $p_{kj} = \widehat{p}_{kj}$ for all j and k. Table 2(b) gives the number of simulated data sets for which $\lambda = 0.01$ produced the smallest value of the criterion. For each of the 3 parameterizations and 5 ICs, estimates from the penalty producing the lowest IC were retained. Table 2(c) shows the mean evaluation criterion averaged over all simulated data sets within each parameterization and IC; standard errors are in parentheses.
*Using the Marker parameterization, 1 Case I data set and 2 Case II data sets could not be fit.

Table 3: Summary of Penalty Search for LOH Application

			Penalties		Robust		Naive	
		κ	α_1	α_2	D.F.	BIC	D.F.	BIC
Ridge Penalty	η independent of group	1	∞	∞	1.0	1757.1	1.0	1757.1
		2	2.5×10^{0}	8.6×10^{0}	11.9	1209.7	11.4	1207.5
		3	2.3×10^{-9}	5.3×10^{0}	18.4	1205.2	21.4	1218.8
		4	2.6×10^{-1}	9.9×10^{4}	13.3	1217.6	13.8	1219.9
	η dependent on group	2	2.5×10^{0}	8.7×10^{0}	12.8	1207.1	12.4	1205.0
		3	2.0×10^{-9}	5.4×10^{0}	20.5	1207.0	23.4	1220.4
		4	∞	∞	–	–	–	–
Lasso Penalty	η independent of group	1	∞	∞	1.0	1757.1	1.0	1757.1
		2	2.2×10^{-1}	2.4×10^{2}	9.1	1215.9	8.0	1211.1
		3	2.0×10^{-1}	8.1×10^{1}	10.7	1206.2	11.0	1207.6
		4	7.4×10^{-1}	1.4×10^{1}	11.5	1208.3	12.0	1210.7
	η dependent on group	2	2.0×10^{-1}	2.5×10^{2}	10.0	1213.4	9.0	1208.7
		3	1.9×10^{-1}	4.8×10^{3}	12.8	1208.2	13.0	1209.0
		4	∞	∞	–	–	–	–

Penalized latent class models applied to the brain tumor data described in the Introduction. Ridge and lasso penalties were used with $\kappa = 1$, 2, 3, and 4 classes. For $\kappa = 2$ and 3, models with homogeneous class prevalence η and with η dependent on MGH/HFH grouping were fit. For the $\kappa = 4$ model with homogeneous η, only the intercept model could produce a positive definite penalized Hessian H_C, and for the group-dependent η even the intercept-only model had a penalized Hessian H_C that was not positive definite.

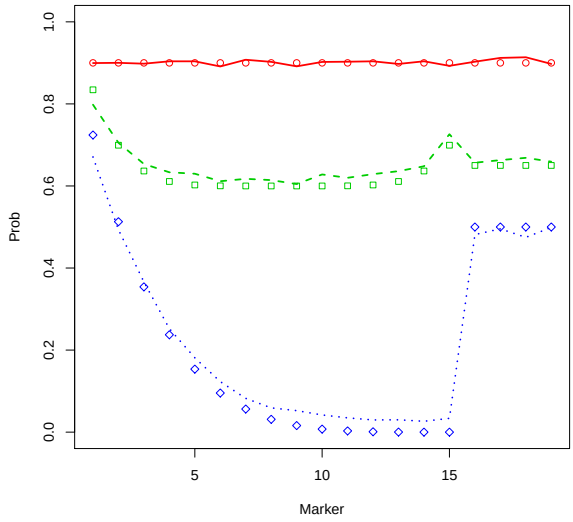
Table 4: Survival Analysis for LOH Classes

Class		Coefficient		Hazard Ratio	
Referent	Comparison	Est	SE	Est	95% Conf. Limit
1	2	-0.67	0.42	0.51	(0.23, 1.2)
1	3	-1.77	0.52	0.17	(0.06, 0.47)
2	3	-1.11	0.45	0.33	(0.14, 0.80)

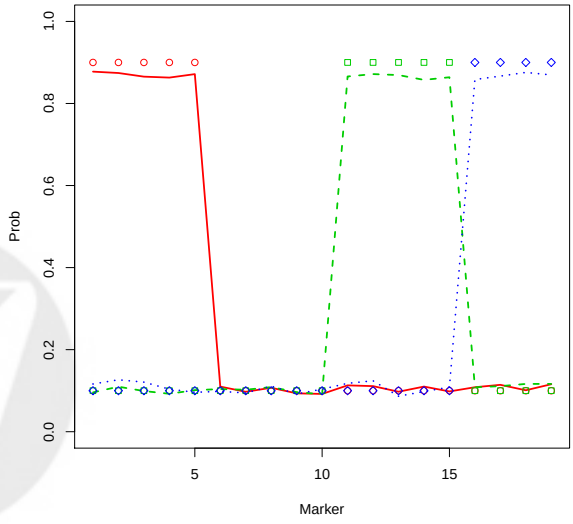
Results from a weighted, stratified Cox proportional hazards model applied to the LOH data described in Sections 1 and 8. Three rows were included for each subject, one for each class, and weighted with the corresponding posterior probability that the subject was a member of the class. Class profiles are depicted in Figure 2. Additionally, subjects were stratified by MGH/HFH group indicator. In the Cox model, class membership comprised the covariates. In the table, "Referent" refers to the reference class and "Comparison" refers to the coefficients and hazard ratios comparing the indicated class against the reference class.

Figure 1: Simulation Results

(a) Case I

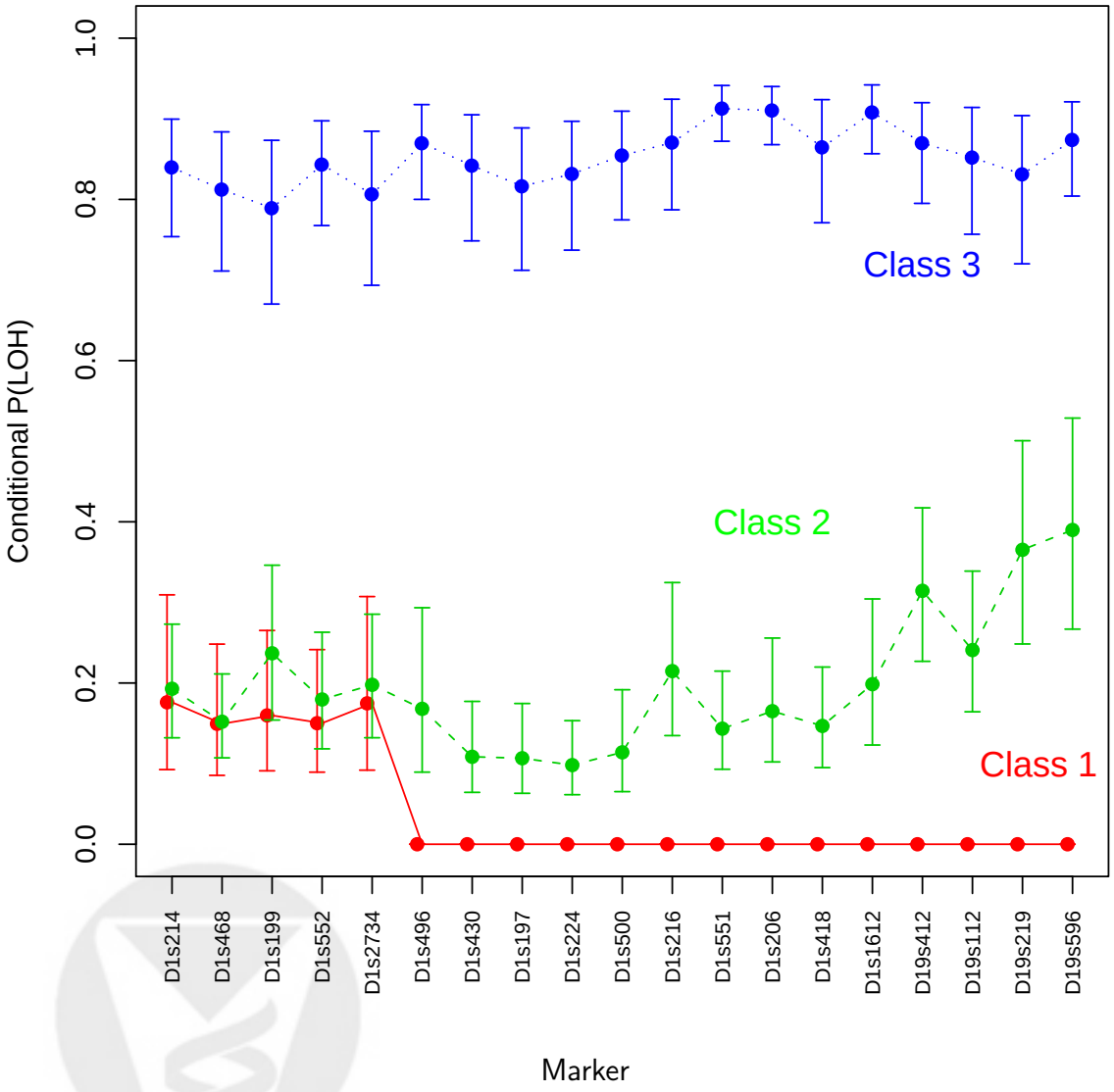


(b) Case II



Symbols represented true probabilities. Lines represent the mean over 250 simulated data sets of estimates obtained from analyses using the Feature I parameterization and the penalty producing the smallest AIC.

Figure 2: Conditional marker probabilities for LOH among Oligodendroglioma patients – best fit with ridge penalty

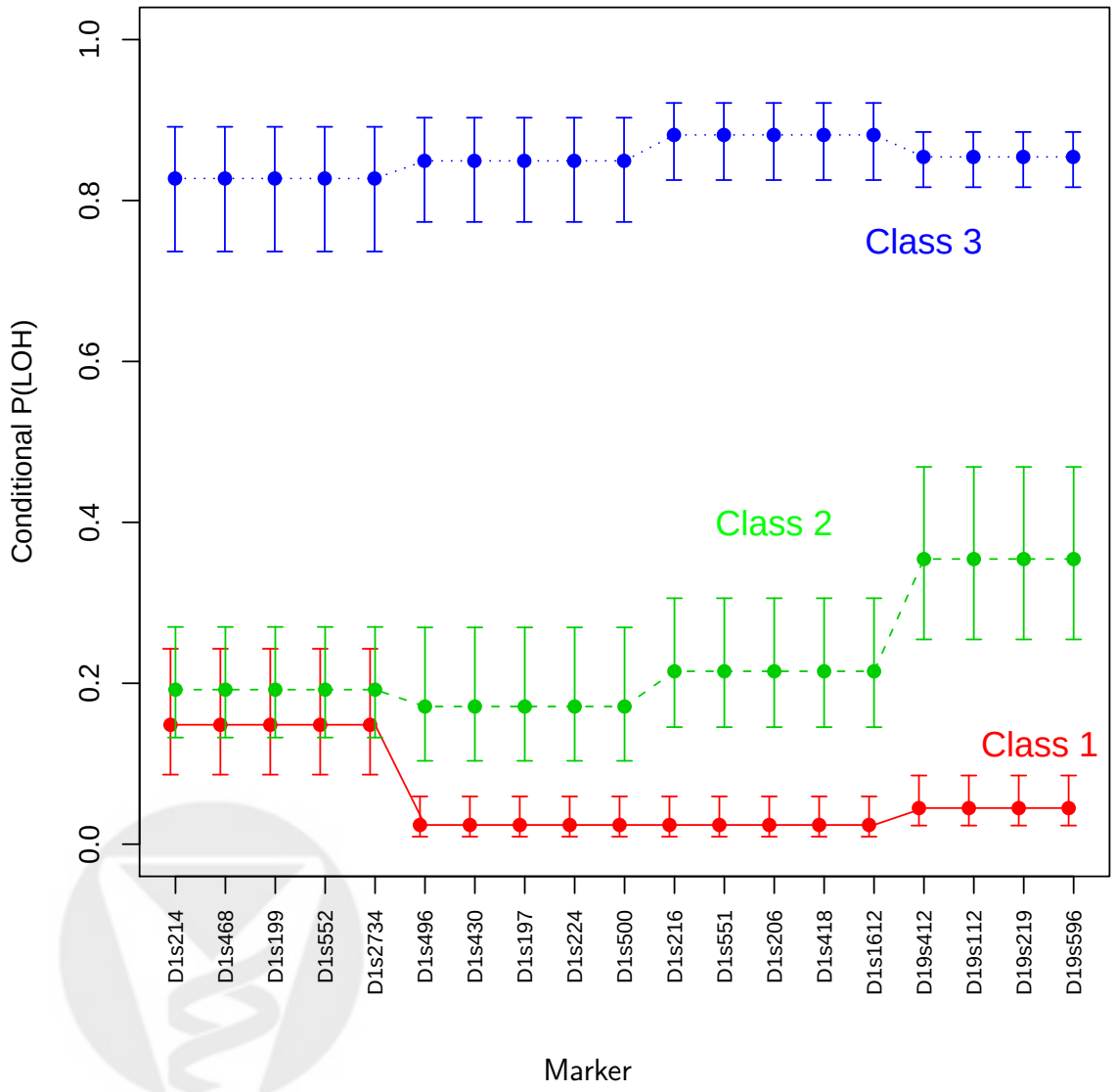


LOH probabilities for each of 19 markers as computed from the combined MGH and HFH data set. Probabilities are based on a ridge penalty. Vertical bars indicate 95% confidence limits based on robust standard errors, conditional on the penalty parameters and the choice $\kappa = 3$.

Legend, with class prevalences by group and corresponding 95% confidence intervals in parentheses:

——	Class 1,	$\eta_{1,MGH} = 0.12\ (0.05, 0.23)$	$\eta_{1,HFH} = 0.25\ (0.14, 0.39)$
- - -	Class 2,	$\eta_{2,MGH} = 0.33\ (0.22, 0.43)$	$\eta_{2,HFH} = 0.48\ (0.40, 0.48)$
.....	Class 3,	$\eta_{3,MGH} = 0.55\ (0.34, 0.73)$	$\eta_{3,HFH} = 0.27\ (0.13, 0.46)$

Figure 3: Conditional marker probabilities for LOH among Oligodendroglioma patients – best fit with lasso penalty



LOH probabilities for each of 19 markers as computed from the combined MGH and HFH data set. Probabilities are based on a lasso penalty. Vertical bars indicate 95% confidence limits based on robust standard errors, conditional on the penalty parameters and the choice $\kappa = 3$.

Legend: —— Class 1; - - - Class 2; ..... Class 3